Hauptseminar
Wintersemester 2012

Corpus Annotation of Information Structure

Abstract:

Information structure has become a central research topics in linguistics for work exploring the interfaces of prosody, syntax, semantics, pragmatics. In computational linguists the analysis of information structure is of equal importance for a range of applications, such as generation and speech synthesis, or the automatic analysis of answers in intelligent language tutoring systems. In support of this theoretical and computational linguistic research, in recent years several research groups have proposed corpus annotation schemes designed to capture aspects of information structure.

In this seminar we want to explore the corpus annotation schemes which have been proposed for information structure and investigate which information structural primitives such as focus/background, given/new can reliably be annotated given which available evidence in the corpus and its meta-information.

Instructors:

- Kordula De Kuthy
  - *Office:* Room 1.26, Blochbau (Wilhelmstr. 19)
  - *Email:* kdk@sfs.uni-tuebingen.de
  - *Office hours:* Tuesdays 10:00–11:00 (please arrange slot by email beforehand)
- Detmar Meurers
  - *Office:* Room 1.28, Blochbau (Wilhelmstr. 19)
  - *Email:* dm@sfs.uni-tuebingen.de
  - *Office hours:* Wednesdays 10:00–11:00 (please arrange slot by email beforehand)

Course meets:

- Wednesdays, 8:30–10:00 in 1.13 (SfS, Blochbau, Wilhelmstr. 19)
- Fridays, 8:30–10:00 in 1.13 (SfS, Blochbau, Wilhelmstr. 19)
  - Note: Following the standard rules, missing more than two meetings unexcused, automatically results in failing the class.

Language:

- The course language is English, but may be switched to German if desired by all.

Credits: 10 CP in MA ISCL

Moodle page: `https://moodle02.zdv.uni-tuebingen.de/course/view.php?id=259`

Syllabus (this file):

- html-Version (http://purl.org/dm/12/ws/infostruc)
- pdf-Version (http://purl.org/dm/12/ws/infostruc/syllabus.pdf)

Nature of course and our expectations: This Hauptseminar intends to provide an overview of the concepts and issues involved in research on information structure and its corpus annotation. Participants are expected to

1. regularly and actively participate in class, read the papers assigned by any of the presenters and post a question on Moodle to the"Reading Discussion Forum" on each reading *at the latest on the day before it is discussed* in class. (20% of grade)
2. explore and present a topic (40% of grade):
   - select one of the sub-topics
   - thoroughly research the topic, taking our literature pointers as a starting point
   - prepare the presentation with slides and discuss the presentation with one of the instructors in the week before the presentation
   - start a new Moodle thread on the "Reading Discussion Forum" specifying what every course participant should read to prepare for your presentation a week before your presentation
   - present the topic in class
3. write and submit a term paper (40% of grade)

Academic conduct and misconduct: Research is driven by discussion and free exchange of ideas, motivations, and perspectives. So you are encouraged to work in groups, discuss, and exchange ideas. At the same time, the foundation of the free exchange of ideas is that everyone is open about where they obtained which information. Concretely, this means you are expected to always make explicit when you've worked on something as a team – and keep in mind that being part of a team always means sharing the work.

For text you write, you always have to provide explicit references for any ideas or passages you reuse from somewhere else. Note that this includes text "found" on the web, where you should cite the url of the web site in case no more official publication is available.

Sessions:

- Week 1:
  - October 17/19: Organization and syllabus
- Week 2
  - October 24: Introduction to Information Structure (Kordula De Kuthy)
    - Reading assignment: Krifka (2007)
  - October 26: Motivating the need for information structure annotation - some empirical evidence (Kordula De Kuthy and Detmar Meurers)
    - Reading assignment: De Kuthy & Meurers (2012)
- Week 3
  - October 31: On Corpus Annotation and Theoretical Linguistics (Detmar Meurers)
    - Reading assignment: Meurers (2005); Meurers & Müller (2009)
  - November 2: The issues as they arise in actual data – hands-on annotation of a small sample corpus
    - Reading Assignment: Dipper et al. (2007)
- Week 4 (November 7, 9):
  - November 7: Background for The English Switchboard Corpus and its framework for annotating information structure in discourse
    - Reading assignment: Prince (1981)
    - Presentation: Emma Li
  - November 9: The English Switchboard Corpus and its framework for annotating information structure in discourse
    - Reading assignment: Calhoun et al. (2010)
      (additional articles of relevance: Nissim et al. (2004), Calhoun et al. (2005))
    - Presentation: Arsenyi Mstislavski
- Week 5 (November 14, 16):
  - The MULI Project. Annotation and Analysis of Information Structure in German and English
    - Reading assignment: Baumann et al. (2004b), Baumann et al. (2004a), Kruijff-Korbayová & Kruijff (2004)
    - Presentation: Anuschka Kranz, Marisa Delz
- Week 6 (November 21, 23):
  - ANNIS: A Linguistic Database for Exploring Information Structure
    - Reading assignment: Ritz et al. (2008), Dipper et al. (2004)

    - Presentation: Daniil Sorokin, Caroline Wagenaar

- Week 7 (November 28, 30):
  - Building and using a richly annotated interlinear diachronic corpus
    - Reading assignment: Donhauser (2007), Petrova et al. (2009)
    - Presentation: Cornelius Fath, Spyridoula Georgatou
- Week 8 (December 5, 7):
  - Annotating Information Status in Spontaneous Speech.
    - Reading assignment: Riester & Baumann (2011), Riester et al. (2010), Baumann & Riester (2010), Baumann & Riester (2012)
    - Presentation: Bettina Remmele, Maria Oberwegner, Velislava Todorova
- Week 9 (December 12, 14):
  - Annotating Information Structure: The Case of Topic
    - Reading assignment: Cook & Bildhauer (2011), Jacobs (2001)
    - Presentation: Tobias Kolditz, Mike Burkhardt
- Week 10 (December 19, 21):
  - Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure (Ramon Ziai)
    - Reading assignment: Meurers, Ziai, Ott & Kopp (2011)
- Week 11 (January 9, 11):
  - Information Structure in African Languages: Corpora and Tools
    - Reading assignment: Chiarcos et al. (2009)
    - Presentation: Marion Zepf, Michael Hahn
- Week 12 (January 16, 18):
  - Annotating Information Structure in a Corpus of Spoken Danish
    - Reading assignment: Paggio (2006)
    - Presentation: Vladlena Sergeeva
- Week 13 (January 23, 25):
  - Tagging of very large corpora in Czech: topic-focus articulation
    - Reading assignment: Buráňová et al. (2000), Hajičová & Sgall (2001), Hajičová et al. (2000), Postolache et al. (2005)
    - Presentation: Maria Chinkina, Maja Bohnacker
  - Annotators Agreement and Evaluation
    - Reading assignment: Vesela et al. (2004), Hajicová et al. (2002)
    - Presentation:
- Week 14 (January 30, February 1): Summary and discussion
- Week 15 (February 6, 8): Presentation of student term paper ideas

References

Baumann, S., C. Brinckmann, S. Hansen-Schirra, G.-J. Kruijff, I. Kruijff-Korbayová, S. Neumann & E. Teich (2004a). Multi-Dimensional Annotation of Linguistic Corpora for Investigating Information Structures. In *Proceedings of NAACL/HLT 2004 Conference Workshop Frontiers in Corpus Annotation*. Boston, MA. URL `acl.ldc.upenn.edu/W/W04/W04-2707.pdf`.

Baumann, S., C. Brinckmann et al. (2004b). The MULI Project. Annotation and Analysis of Information Structure in German and English. In *Proceedings of LREC 2004*. Lisbon. URL `www.coli.uni-saarland.de/~cabr/papers/muli.LREC2004main.pdf`.

Baumann, S. & A. Riester (2010). Annotating Information Status in Spontaneous Speech. In *Proceedings of the Fifth International Conference on Speech Prosody*. Chicago. URL `http://www.ims.uni-stuttgart.de/~arndt/doc/baumannRiesterSpeechPros2010`.

Baumann, S. & A. Riester (2012). Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects. In G. Elordieta & P. Prieto (eds.), *Prosody and Meaning*, Berlin: Mouton de Gruyter, vol. 25 of *Interface Explorations*. URL `http://www.ims.uni-stuttgart.de/~arndt/doc/baumannRiesterBarcelonaPrefinal.pdf`.

Buráňová, E., E. Hajičová & P. Sgall (2000). Tagging of very large corpora: topic-focus articulation. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, COLING '00, pp. 139–144. URL `http://aclweb.org/anthology-new/C/C00/C00-1021.pdf`.

Calhoun, S., J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman & D. Beaver (2010). The NXT-format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. *Language Resources and Evaluation* 44, 387–419.

Calhoun, S., M. Nissim, M. Steedman & J. Brenier (2005). A Framework for Annotating Information Structure in Discourse. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 45–52. URL `http://aclweb.org/anthology/W/W05/W05-0307`.

Chiarcos, C., I. Fiedler et al. (2009). Information Structure in African Languages: Corpora and Tools. In *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009*. Athens, Greece, pp. 17 – 24. URL `http://aflat.org/files/W09-0703.pdf`.

Cook, P. & F. Bildhauer (2011). Annotating information structure. The case of "topic". In S. Dipper & H. Zinsmeister (eds.), *Beyond Semantics. Corpus based Investigations of Pragmatic and Discourse Phenomena*, Ruhr Universität Bochum, Bochumer Linguistische Arbeitsberichte, p. 45–56. URL `http://www.linguistics.ruhr-uni-bochum.de/bla/beyondsem2011/cook_final.pdf`.

De Kuthy, K. & D. Meurers (2012). Focus projection between theory and evidence. In S. Featherston & B. Stolterfoth (eds.), *Empirical Approaches to Linguistic Theory – Studies in Meaning and Structure*, De Gruyter, vol. 111 of *Studies in Generative Grammar*, pp. 207–240. URL `http://purl.org/dm/papers/dekuthy-meurers-11.html`.

Dipper, S., M. Götze & S. Skopeteas (eds.) (2007). *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*, vol. 7 of *Interdisciplinary Studies on Information*

Structure. Potsdam, Germany: Universitätsverlag Potsdam. URL http://www.sfb632.uni-potsdam.de/publications/isis07.pdf.

Dipper, S., M. Götze, M. Stede & T. Wegst (2004). ANNIS: A Linguistic Database for Exploring Information Structure. In S. Ishihara, M. Schmitz & A. Schwarz (eds.), *Interdisciplinary Studies on Information Structure*, Potsdam: University publishing house Potsdam, vol. 1 of *Working Papers of the SFB632*, pp. 245 – 279. URL http://www.sfb632.uni-potsdam.de/publications/isis01_7dipper-etal.pdf.

Donhauser, K. (2007). Zur informationsstrukturellen Annotation sprachhistorischer Texten. *Sprache und Informationsverarbeitung* 31, 39–45. URL http://www.sfb632.uni-potsdam.de/publications/B4/donhauser_2007.pdf.

Hajicová, E., P. Pajas & K. Vesela (2002). Corpus Annotation on the Tectogrammatical Layer: Summarizing of the First Stages of Evaluations. *Prague Bull. Math. Linguistics* 77, 5–18. URL http://ufal.mff.cuni.cz/pbml/77/hajicova-et-al.pdf.

Hajičová, E. & P. Sgall (2001). Topic-focus and salience. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Toulouse, France: Association for Computational Linguistics, ACL '01, pp. 276–281. URL http://aclweb.org/anthology-new/P/P01/P01-1036.pdf.

Hajičová, E., J. Panevová & P. Sgall (2000). *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*. Technical report tr-2000-09, ÚFAL/CKL. URL http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/tmanual/tmanen.pdf. In cooperation with A. Böhmová, M. Ceplová and V. Řezníčková Translated by Z. Kirschner, E. Hajičová and P. Sgall.

Jacobs, J. (2001). The dimensions of topic-comment. *Linguistics* 39, 641–681. URL http://www.degruyter.com/dg/viewarticle.fullcontentlink:pdfeventlink/contentUri?format=INT&t:ac=j$002fling.2001.39.issue-4$002fling.2001.027$002fling.2001.027.xml.

Krifka, M. (2007). Basic Notions of Information Structure. In C. Fery, G. Fanselow & M. Krifka (eds.), *The notions of information structure*, Potsdam: Universitätsverlag Potsdam, vol. 6 of *Interdisciplinary Studies on Information Structure (ISIS)*. URL http://www.sfb632.uni-potsdam.de/publications/isis06_2krifka.pdf.

Kruijff-Korbayová, I. & G.-J. M. Kruijff (2004). Discourse-level Annotation for Investigating Information Structure. In B. Webber & D. K. Byron (eds.), *ACL 2004 Workshop on Discourse Annotation*. Barcelona, Spain: Association for Computational Linguistics, pp. 41–48. URL http://www.aclweb.org/anthology-new/W/W04/W04-0206.pdf.

Meurers, D., R. Ziai, N. Ott & J. Kopp (2011). Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*. Edinburgh, Scotland, UK: Association for Computational Linguistics, pp. 1–9. URL http://aclweb.org/anthology/W11-2401.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115(11), 1619–1639. URL http://purl.org/dm/papers/meurers-03.html.

Meurers, W. D. & S. Müller (2009). Corpora and Syntax (Article 42). In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics*, Berlin: Mouton de Gruyter, vol. 2 of *Handbooks of Linguistics and Communication Science*, pp. 920–933. URL http://purl.org/dm/papers/meurers-mueller-09.html.

Nissim, M., S. Dingare, J. Carletta & M. Steedman (2004). An annotation scheme for information status in dialogue. In *Proceedings of the 4th Conference on Language Resources and Evaluation*. Lisbon, Portugal. URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/638.pdf.

Paggio, P. (2006). Annotating Information Structure in a Corpus of Spoken Danish. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy, pp. 1606 – 1609. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/639_pdf.pdf.

Petrova, S., M. Solf, J. Ritz, C. Chiarcos & A. Zeldes (2009). Building and using a richly annotated interlinear diachronic corpus: the case of Old High German Tatian. *Traitement Automatique des Langues* 50(2), 47–71. URL http://www.sfb632.uni-potsdam.de/publications/B4/Petrova_et_al_2009_tal.pdf.

Postolache, O., I. Kruijff-Korbayova & G.-J. Kruijff (2005). Data-driven Approaches for Information Structure Identification. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics. URL http://www.aclweb.org/anthology/H/H05/H05-1002.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (ed.), *Radical Pragmatics*, New York: Academic Press, p. 223–256. URL http://www.ling.upenn.edu/~ellen/givennew.pdf.

Riester, A. & S. Baumann (2011). Information Structure Annotation and Secondary Accents. In S. Dipper & H. Zinsmeister (eds.), *Beyond Semantics. Corpus-based Investigations of Pragmatic and Discourse Phenomena*, Ruhr Universität Bochum, vol. 3 of *Bochumer Linguistische Arbeitsberichte*, pp. 111–127. URL http://www.ims.uni-stuttgart.de/~arndt/doc/baumannRiesterBeyondSem.pdf.

Riester, A., D. Lorenz & N. Seemann (2010). A Recursive Annotation Scheme for Referential Information Status. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/764_Paper.pdf.

Ritz, J., S. Dipper & M. Götze (2008). Annotation of Information Structure: An Evaluation Across Different Types of Texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco, pp. 2137–2142. URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/543_paper.pdf.

Vesela, K., J. Havelka & E. Hajicová (2004). Annotators Agreement: The Case of Topic-Focus Articulation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. European Language Resources Association. URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/350.pdf.

Last update: November 7, 2012